Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases
of Complex and Scanned Documents

ARPA Order No. D570

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: July 14, 1999

Period of Report: April 1, 1999 to June 30, 1999

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	07/14/99	Scientific/Tech

4. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents

5. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D570

6. AUTHOR(S)
W. Bruce Croft

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

8. PERFORMING ORGANIZATION REPORT NUMBER
TR5281810799

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Mr. Harry Koch
ESC/AXS
Bldg 1704. Room 114
5 Eglin St.
Hanscom AFB, MA 01731-2116

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

14. SUBJECT TERMS
Browsing, Query Processing, Indexing, Image Retrieval, Scanned Document Retrieval, Bayesian Network, Text Retrieval, Probabilistic Retrieval Model, Large Distributed Databases

15. NUMBER OF PAGES
9

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we have been studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, extending the underlying retrieval model to be able to make effective use of phrasal representations, and using complex document structure in indexing and retrieval.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. As well as the PTO text databases, we will make extensive use of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query.

Technical Results

We have developed a new retrieval based on language models. One advantage of language models is that they provide a better theoretical foundation for retrieval. This approach has showed a lot of promise with some data – in fact initial experiments with a crude version of the model have shown that it can perform as well as other systems at TREC. We are now carrying out further experiments to see the effects of learned language models on retrieval.

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

Language models may provide retrieval improvements for PTO data.

Task 2: Browsing and Classification Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing and classifying collections of documents. These techniques will be designed to support interactive browsing and text classification in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. The classification task emphasizes the ability to accurately assign predefined categories (as in the PTO classification) to new documents (patents). An additional problem is to determine when existing classifications do not match well to new documents, such as when a PTO category covers too many patents and needs to be refined.

General Methodology

Evaluation of these techniques will be done using both the TREC corpus and PTO data. For the classification task in particular, we are designing evaluation criteria with substantial input from PTO staff.

Technical Results

We are working on a multi-level scheme for classification which involves dividing a database into smaller databases by class and then using collection selection to find the appropriate class. A number of different approaches to collection selection are being

investigated. We have been able to utilize some of our results on distributed retrieval for this problem.

As a result of the last PTO meeting in May, we are focusing our work on evaluation and technology transfer of our results. One consequence of this is that we have defined a project with Dataware to evaluate different approaches to hierarchical classification.

In the summarization/visualization area we have developed a system for combining a ranked list with clustering. A ranked list is a well-known technique for presenting information so that relevant documents may be found quickly. Clustering is also a well-known technique for grouping similar documents. By combining the two, we have developed an approach that exceeds the retrieval effectiveness of a traditional ranked list. This work is described in the following new papers:

- Leuski, A. and Allan, J., "The Best of Both Worlds: Combining Ranked List and Clustering," submitted to the *Eighth International Conference on Information and Knowledge Management, (CIKM99)*, Kansas City, MO, November 2-6, 1999.
- Leuski, A., "Studying the Usability of Relevant Proximity Ranking," submitted as a poster presentation to 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99), Berkeley, CA, August 15-19, 1999

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We will continue to improve the demonstration system and plan to carry out further classification experiments using the collection mechanism. We will also be doing an evaluation of hierarchical classification with Dataware.

Task 3: Image Indexing and Retrieval

Task Objectives

The goal of this task is to develop similarity-based techniques for retrieving images such as trademarks, logos, and designs.

Technical Problems

The central issue is how images can be indexed to support efficient, content-based retrieval. The primary type of query in these environments is "find me things that look like this". We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with hundreds of thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images. Specifically, we are working to obtain large collections of trademark and design images, both from the PTO and from general sources such as the web.

Technical Results

As a result of the last meeting, our work is now focused on evaluating the demonstration system and on technology transfer. We have received a complete collection of geometric trademarks with relevance judgements from the British Patent Office. The relevance judgements were performed by a trademark examiner. We are currently in the process of indexing this database so that the performance of the demonstration system can be evaluated. We are also creating a user interface for obtaining relevance judgements on the USPTO database so that we may be able to evaluate the effectiveness of the trademark demonstration system on the PTO database. We also continue to improve the effectiveness of our trademark retrieval system. We have collected 4000 additional flower images from the web. We will be indexing this and adding them to the flower patent database to judge retrieval effectiveness over a larger set.

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implications for Further Research

We continue to focus on evaluating the accuracy of our techniques using trademark testbeds from Britain and, hopefully, from the U.S. PTO. We will also continue to improve the demonstration system.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client-server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using TREC data and large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

We are now doing a more controlled study of whether it is better to organize documents chronologically or by subject. Results from other research suggest that it is better to organize by subject. We are verifying with PTO data.

We are nearing completion on a paper for the journal ACM TOIS that will describe the results of the query-sampling work.

Results from distributed retrieval are being used for the classification problem (see Task 2).
More experiments are being carried out with the language model approach to collection selection to evaluate its effectiveness in terms of precision/recall.

We are also doing more experiments to see the effects of resource descriptions on precision/recall in document retrieval. Some of this work is described in the following reports.

- Lu, Z. and McKinley, K., "Searching a Terabyte of Text Using Partial Replication," CIIR Technical Report.
- Lu, Zhihong , "Scalable Distributed Architectures for Information Retrieval," Ph.D. dissertation.

Important Findings and Conclusions

Organization by topic is likely to be better than a chronological organization. Distributed search can be more effective than centralized search if it is based on language models. Replication can significantly improve the performance (response time) of a large distributed system.

Significant Hardware Development

None.

Special Comments

None

Implications for Further Research

Organizing documents by subject is likely to be more effective than organizing them by the date of the document. Appropriate resource selection with improve speed while not reducing effectiveness drastically. We will continue to evaluate performance of distributed architectures for scalable IR using the new demonstration system.